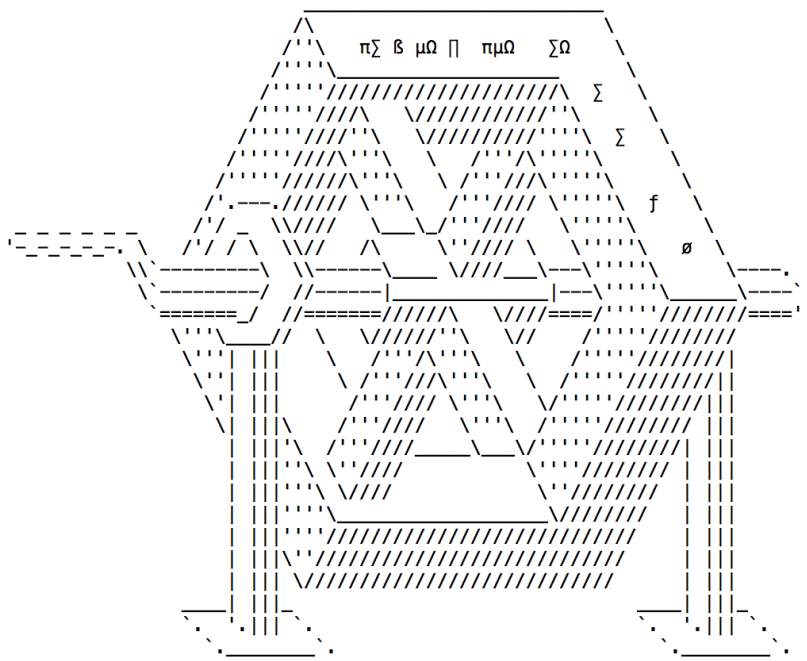


The Matsu Wheel:

0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, ...

A Cloud-based Scanning Framework for Analyzing Large Volumes of Hyperspectral Data

Maria Patterson, PhD
Open Science Data Cloud
Center for Data Intensive Science (CDIS)
University of Chicago

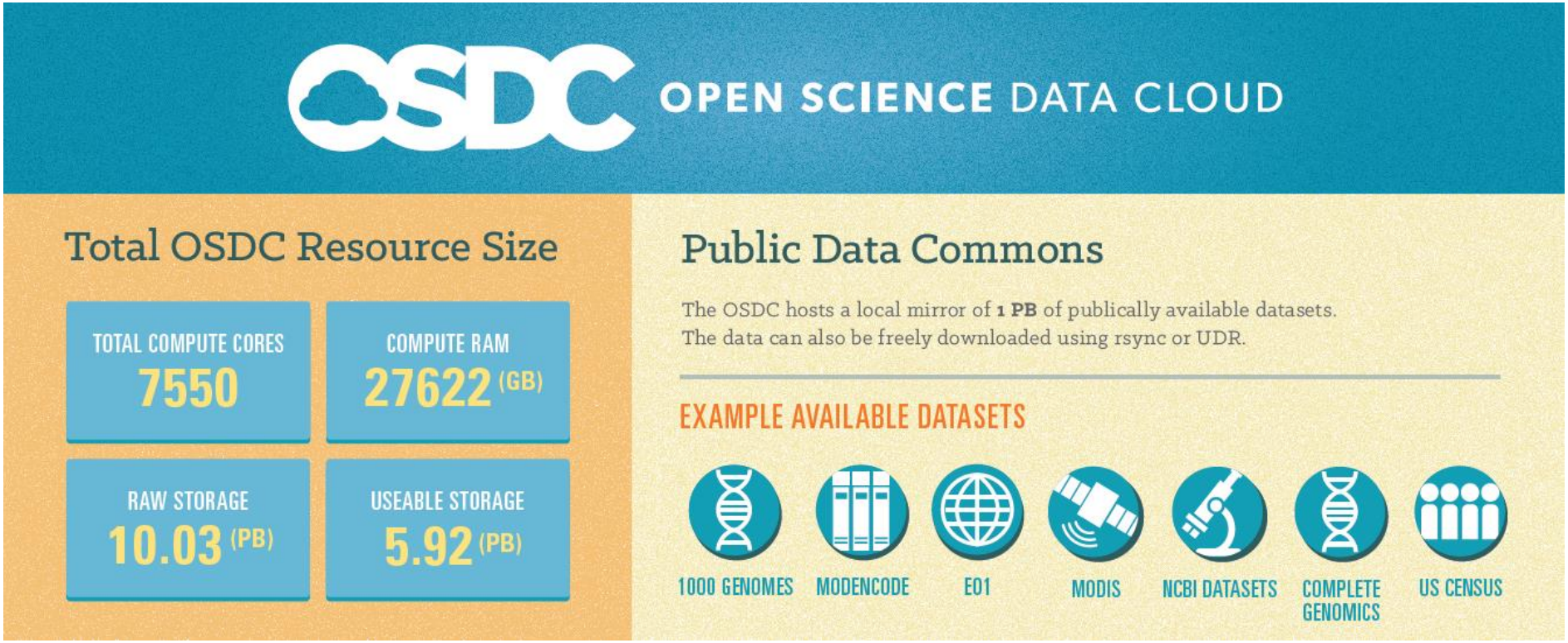
HyspIRI Symposium, 5 June, 2014

THE UNIVERSITY OF CHICAGO

OPEN SCIENCE DATA CLOUD

OPEN CLOUD CONSORTIUM

The Open Science Data Cloud (OSDC) is an **open-source**, **cloud-based** infrastructure that allows scientists to manage, share, and analyze medium to large size scientific datasets.



Application for resources available to anyone doing scientific research:

www.opensciencedatacloud.org

Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*

User view: 1) login

OSDC Console Apply Public Data Systems Projects Status Support News PIRE

OSDC OPEN SCIENCE DATA CLOUD

Cloud Services for the Scientific Community

The OSDC provides petabyte-scale cloud resources that let you easily analyze, manage, and share data.

Get Started Now | OSDC Console Login

Featured on the OSDC

How can I get involved?

Apply

Fill out a short application for an OSDC resource allocation. Allocations start at 16 dedicated cores and 1TB of storage, but scale depending on the project needs and level of organizational partnership.

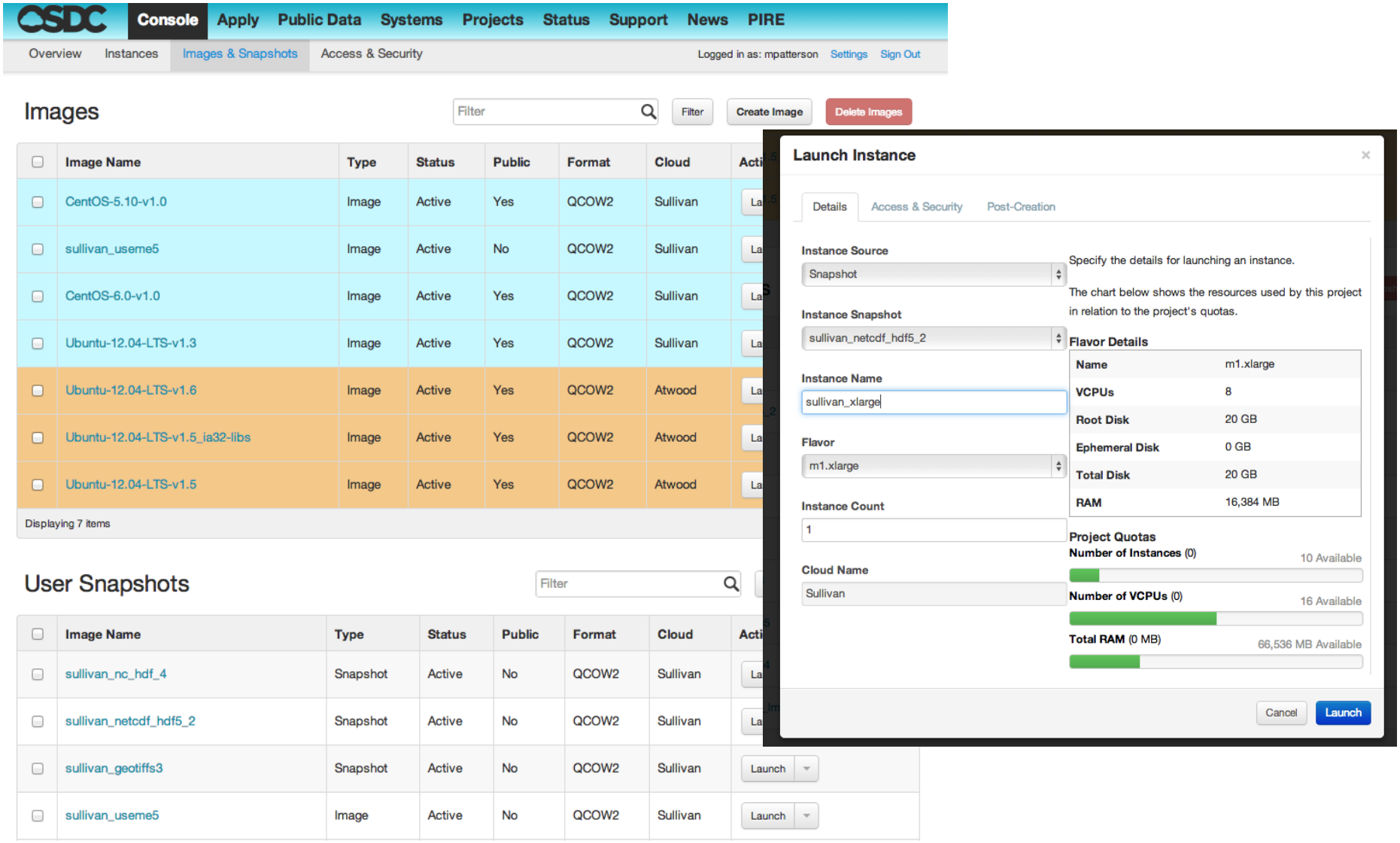
Partner

Partner with us and add your own racks to the OSDC (we will manage them for you). Organizations can also join the Open Cloud Consortium (OCC) which is made up of working groups, including the OSDC.

Develop

All of the software developed as part of the OSDC is open source and hosted on GitHub. You can directly help the scientific cloud computing community by contributing to the open source OSDC software stack.

User view: 2) launch virtual machine



Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*

User view: 3) run analysis

OSDC | Console | Apply | Public Data | Systems | Projects | Status | Support | News | PIRE

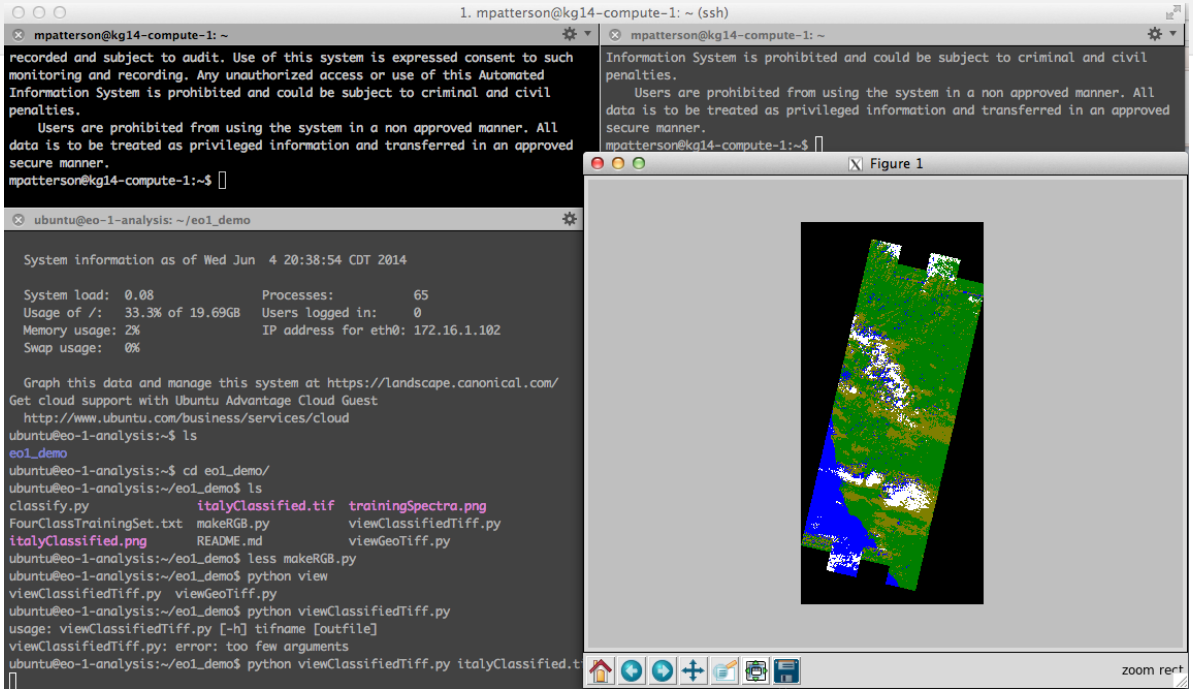
Overview | Instances | Images & Snapshots | Access & Security | Logged in as: mpatterson | Settings | Sign Out

Instances

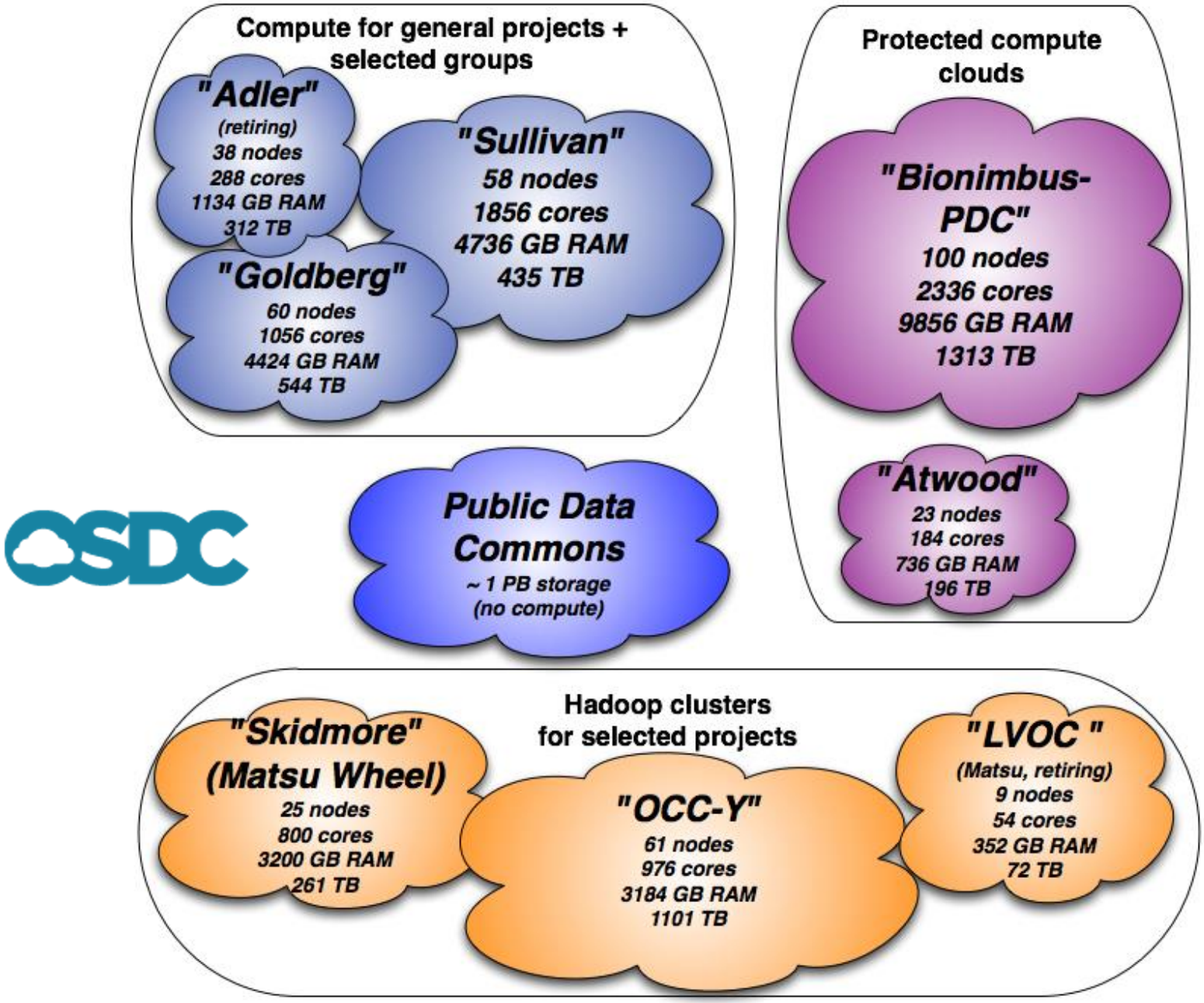
Launch Instance | Terminate Instances

Instance Name	IP Address	Size	Keypair	Status	Task	Power State	Cloud	Actions
EO-1 analysis	172.16.1.102	m1.small | 2GB RAM | 1 VCPU | 20GB Disk	osdc_keypair	Active	None	Running	Sullivan	Create Snapshot

Displaying 1 item



Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*



Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*

Project Matsu

- Joint effort between the Open Cloud Consortium (lead, Robert Grossman) and NASA (lead, Dan Mandl) to develop open source technology for cloud-based processing of satellite imagery to support earth sciences.

- The OSDC is used to process Earth Observing 1 (EO-1) *satellite imagery* from the Advanced Land Imager and the Hyperion instruments and to make this data available to interested users.
  - Namibia flood dashboard, WCPS
  - Hadoop-based ‘Matsu Wheel’ scanning data algorithm

Matsu Analytic Wheel

Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*

Matsu Analytic Wheel

New data observed by EO-1 and downloaded to NASA

NASA Goddard Space Flight Center

NASA images sent to OSDC Public Data Commons cloud for permanent storage

OSDC Public Data Commons (GlusterFS)

- The Wheel "watches" for new data to become available, using Apache Storm.
- When new data are detected, loaded into Hadoop's distributed file system for analysis using MapReduce.
- The Wheel analytics run each night, daily reports available the morning after data are received.

Matsu Analytic Wheel

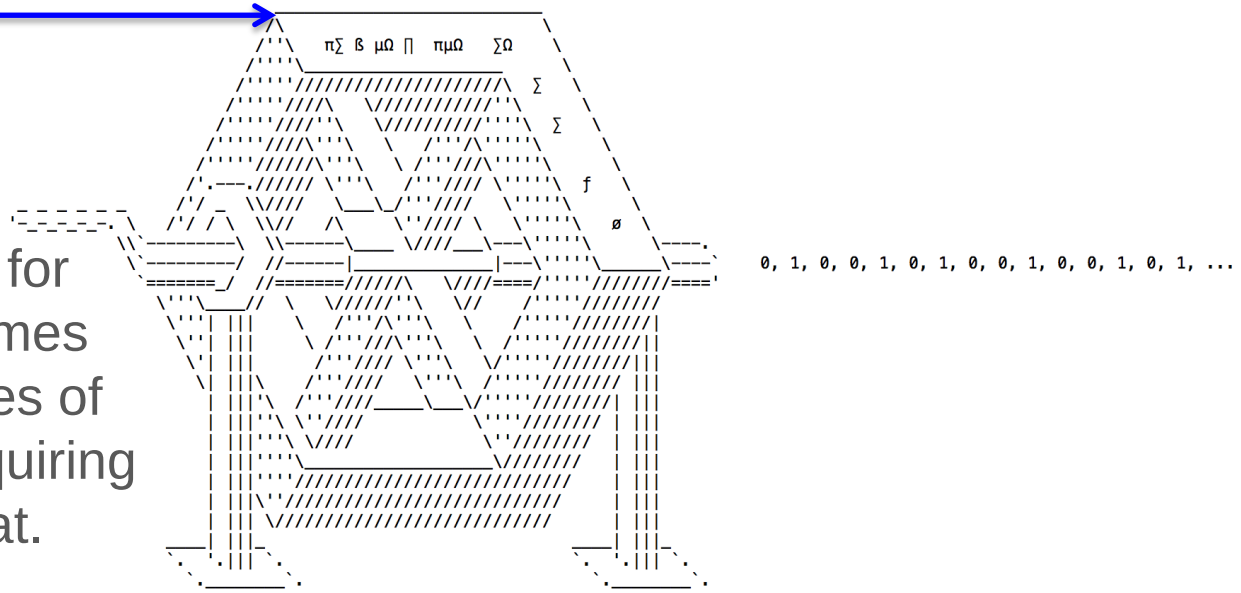
New data observed by EO-1 and downloaded to NASA

NASA Goddard Space Flight Center

NASA images sent to OSDC Public Data Commons cloud for permanent storage

OSDC Public Data Commons (GlusterFS)

Data read into HDFS only once



- The Wheel is efficient for processing large volumes of data with many types of analysis by simply requiring a common input format.

Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*

Matsu Analytic Wheel

Earth Observing-1

New data observed by EO-1 and downloaded to NASA

NASA Goddard Space Flight Center

NASA images sent to OSDC Public Data Commons cloud for permanent storage

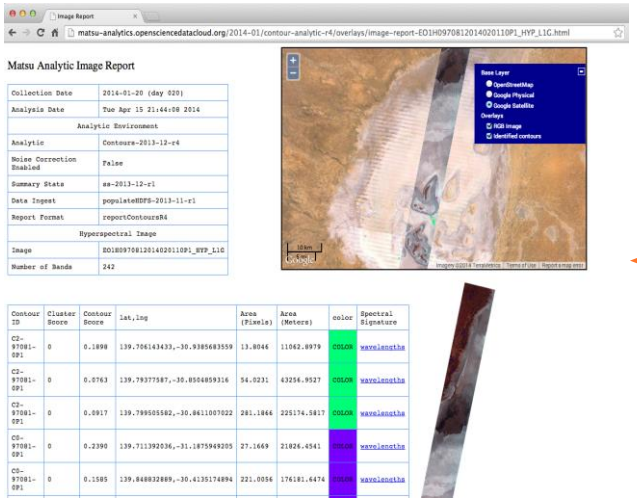
OSDC Public Data Commons (GlusterFS)

Data read into HDFS only once

HDFS

Wheel analytics run over data using MapReduce

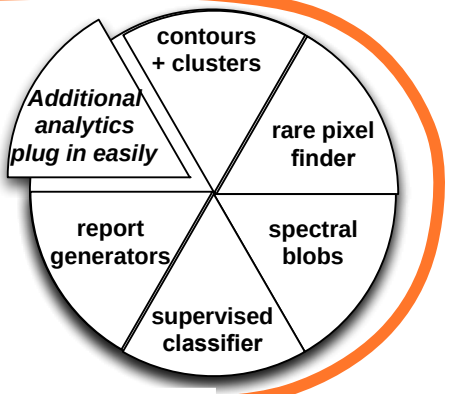
Metadata stored



Analytic reports generated by Wheel are accessible via web browser

NoSql Database (Accumulo)

Analytic results stored

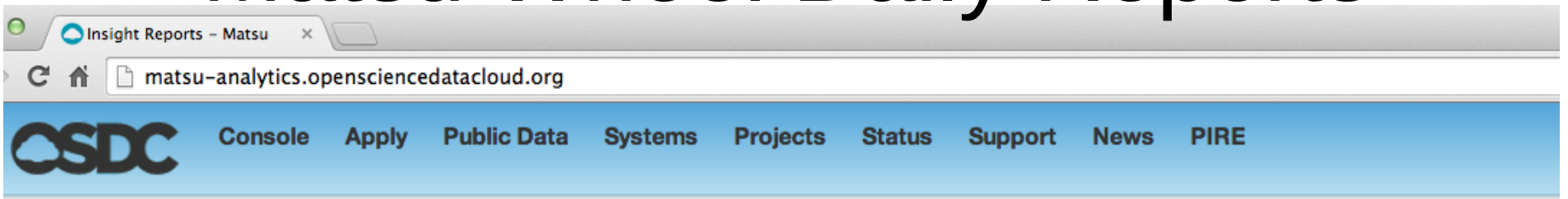


Secondary analysis can be done from analytic database

Maria Patterson (mtpatter@uchicago.edu)

Center for Data Intensive Science, University of Chicago

Matsu Wheel Daily Reports



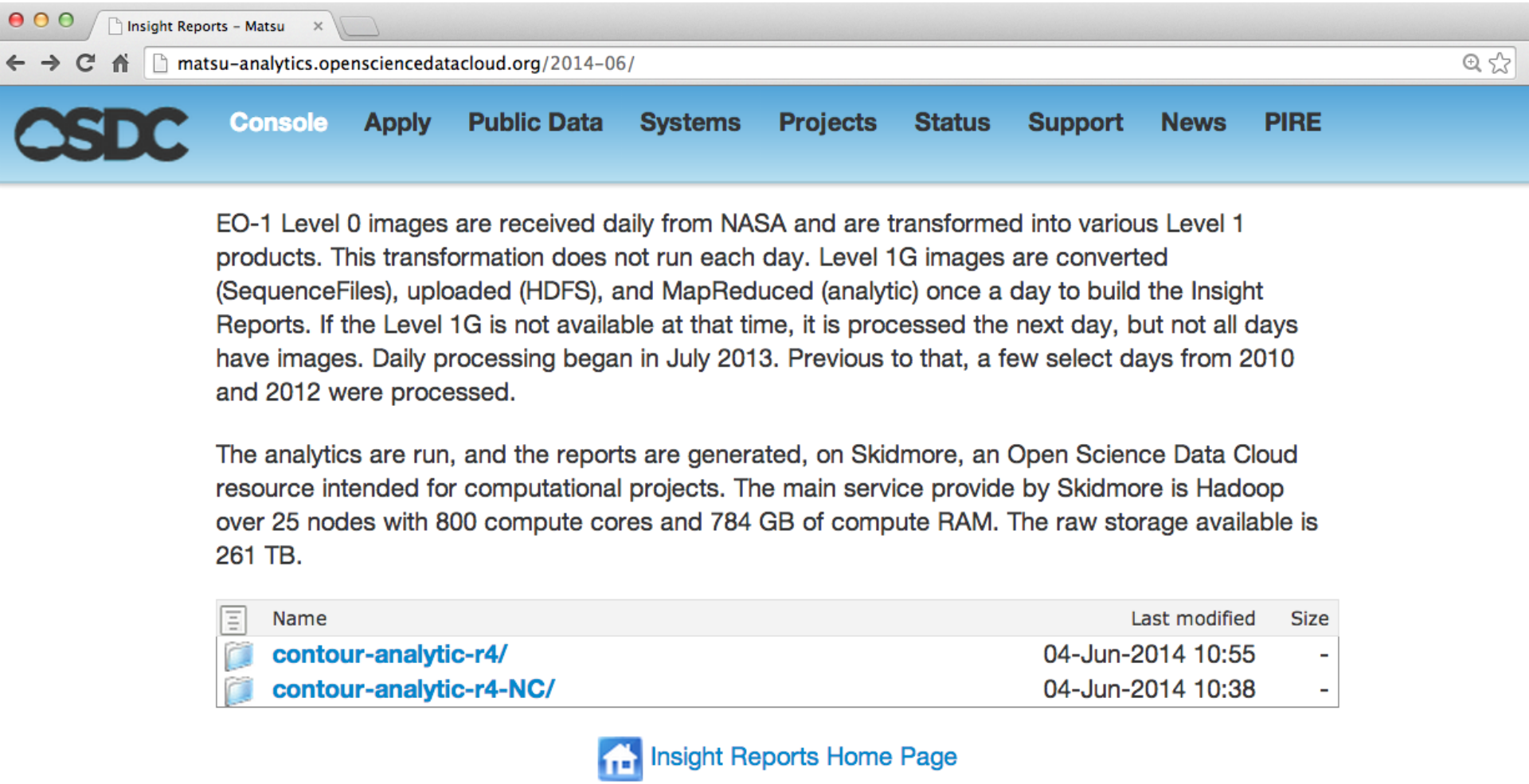
EO-1 Level 0 images are received daily from NASA and are transformed into various Level 1 products. This transformation does not run each day. Level 1G images are converted (SequenceFiles), uploaded (HDFS), and MapReduced (analytic) once a day to build the Insight Reports. If the Level 1G is not available at that time, it is processed the next day, but not all days have images. Daily processing began in July 2013. Previous to that, a few select days from 2010 and 2012 were processed.

The analytics are run, and the reports are generated, on Skidmore, an Open Science Data Cloud resource intended for computational projects. The main service provide by Skidmore is Hadoop over 25 nodes with 800 compute cores and 784 GB of compute RAM. The raw storage available is 261 TB.

Name	Last modified	Size
2014-06/	04-Jun-2014 10:18	-
2014-05/	01-Jun-2014 10:23	-
2014-04/	20-Apr-2014 12:23	-
2014-03/	20-Apr-2014 11:47	-
2014-02/	20-Apr-2014 11:06	-
2014-01/	16-Apr-2014 16:13	-
2013-12/	25-Mar-2014 17:18	-
2013-11/	25-Mar-2014 17:18	-
2013-10/	25-Mar-2014 17:18	-
2013-09/	25-Mar-2014 17:18	-
2013-08/	25-Mar-2014 17:18	-
2013-07/	25-Mar-2014 17:18	-
2012/	25-Mar-2014 17:18	-
2010/	25-Mar-2014 17:18	-

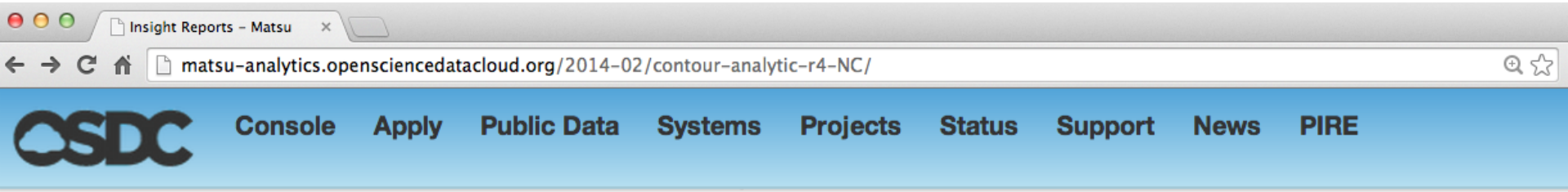
matsu-analytics.opensciencedatacloud.org

Matsu Wheel Daily Reports



Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*

Matsu Wheel Daily Reports



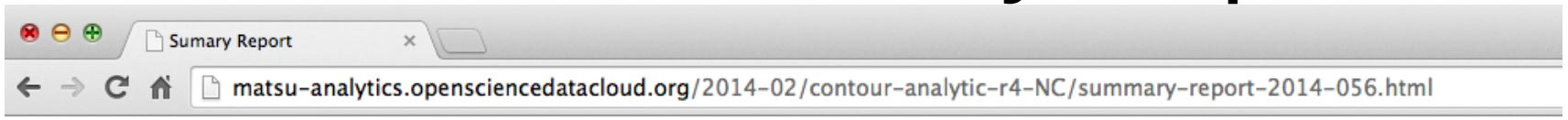
EO-1 Level 0 images are received daily from NASA and are transformed into various Level 1 products. This transformation does not run each day. Level 1G images are converted (SequenceFiles), uploaded (HDFS), and MapReduced (analytic) once a day to build the Insight Reports. If the Level 1G is not available at that time, it is processed the next day, but not all days have images. Daily processing began in July 2013. Previous to that, a few select days from 2010 and 2012 were processed.

The analytics are run, and the reports are generated, on Skidmore, an Open Science Data Cloud resource intended for computational projects. The main service provide by Skidmore is Hadoop over 25 nodes with 800 compute cores and 784 GB of compute RAM. The raw storage available is 261 TB.

Name	Last modified	Size
overlays/	17-Apr-2014 12:34	-
summary-report-2014-059.html	17-Apr-2014 03:12	7.3K
summary-report-2014-058.html	17-Apr-2014 02:43	7.5K
summary-report-2014-057.html	17-Apr-2014 02:14	7.7K
summary-report-2014-056.html	17-Apr-2014 01:45	8.1K
summary-report-2014-054.html	17-Apr-2014 01:10	7.7K

Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*

Matsu Wheel Daily Reports

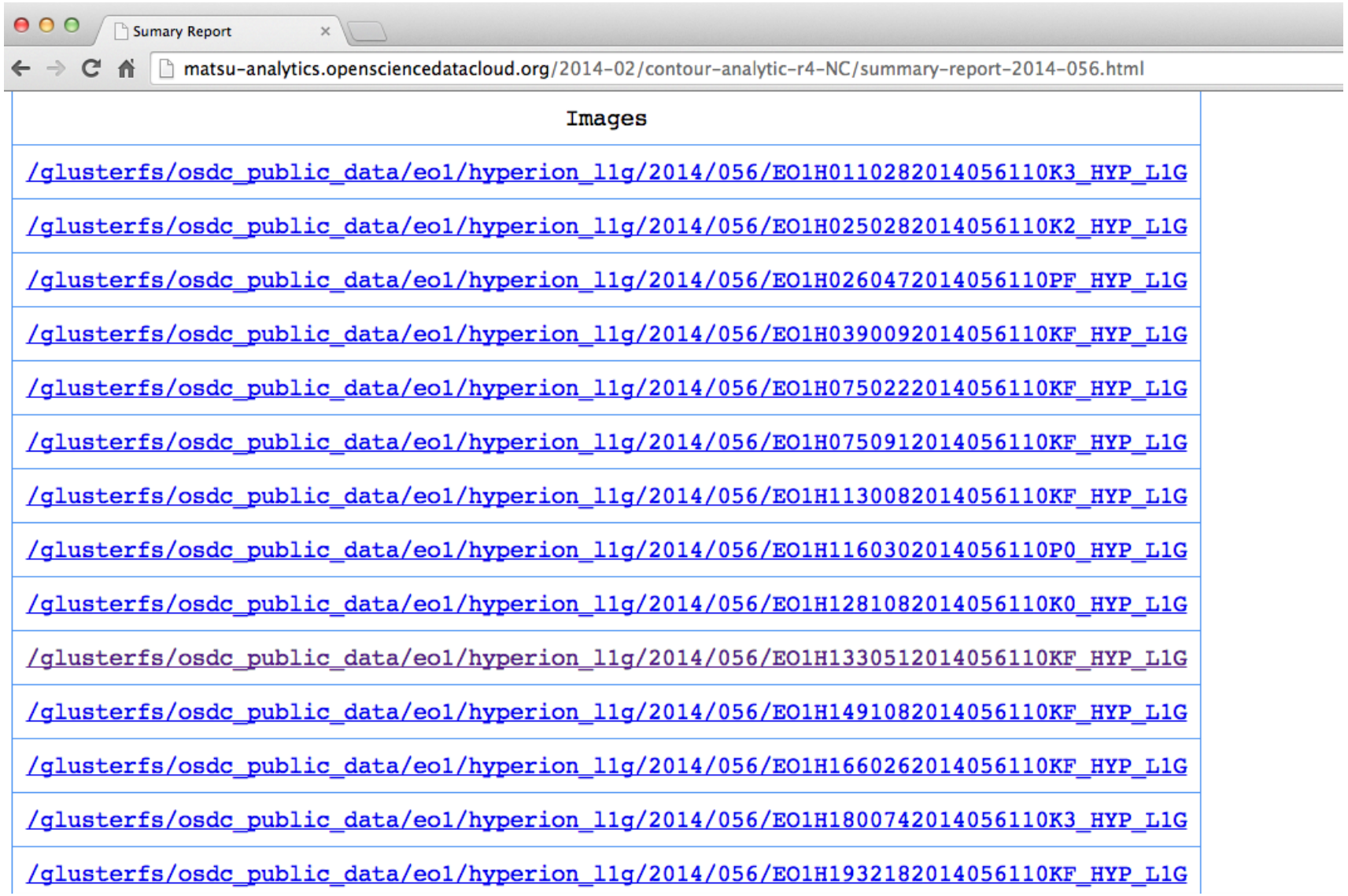


Analytic Summary Report

Collection Date	2014-02-25 (day 056)
Analysis Date	Thu Apr 17 01:45:32 2014
Analytic Environment	
Analytic	Contours-2013-12-r4
Noise Correction Enabled	True
Summary Stats	ss-2013-12-r1
Data Ingest	populateHDFS-2013-11-r1
Report Format	reportContoursR4
Run Summary	
Number of Image	16
Average Number of Pixels	922639
Image with Largest Variance	EO1H0260472014056110PF_HYP_L1G

Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*

Matsu Wheel Daily Reports



Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*

Matsu Wheel Daily Reports

Sumary Report

matsu-analytics.opensciencedatacloud.org/2014-02/contour-analytic-r4-NC/summary-report-2014-056.html

	Hyper-spectral Objects				
Rank	Object Name	Cluster Score	Contour Score	Location (lng/lat)	Image
1	C0-33051-0KF	1000	0.7446	93.8739019579,12.2704698125	/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G
2	C2-33051-0KF	1000	0.6196	93.8682455446,12.2799335922	/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G
3	C2-33051-0KF	1000	0.4510	93.8585620292,12.2742192321	/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G
4	C2-33051-0KF	1000	0.4176	93.8552204394,12.2723263195	/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G
5	C0-33051-0KF	1000	0.3845	93.8670069173,12.2691974168	/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G
6	C3-33051-0KF	1000	0.3512	93.8656741679,12.2854892912	/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G
7	C0-33051-0KF	1000	0.2894	93.8752966894,12.2775235269	/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G
8	C2-33051-0KF	1000	0.2361	93.8653887507,12.2746401012	/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G
9	C1-26086-0K1	1000	0.1580	-61.3432105294,-37.5568381617	/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H2260862014056110K1_HYP_L1G

Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*

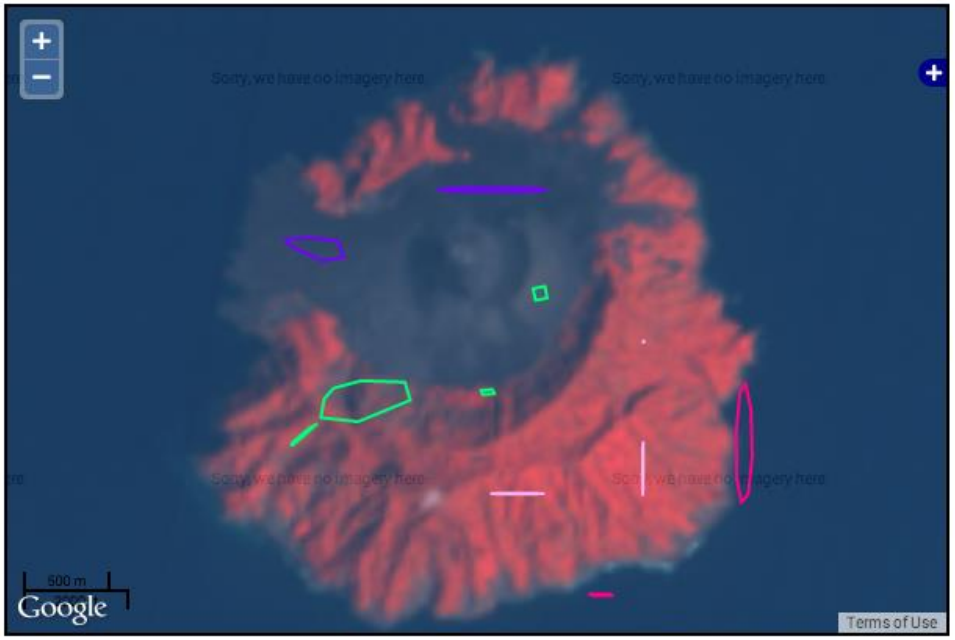
Matsu Wheel Daily Reports

Image Report

matsu-analytics.opensciencedatacloud.org/2014-02/contour-analytic-r4-NC/overlays/image-report-EO1H1330512014056110KF_HYP_L1G.html

Matsu Analytic Image Report

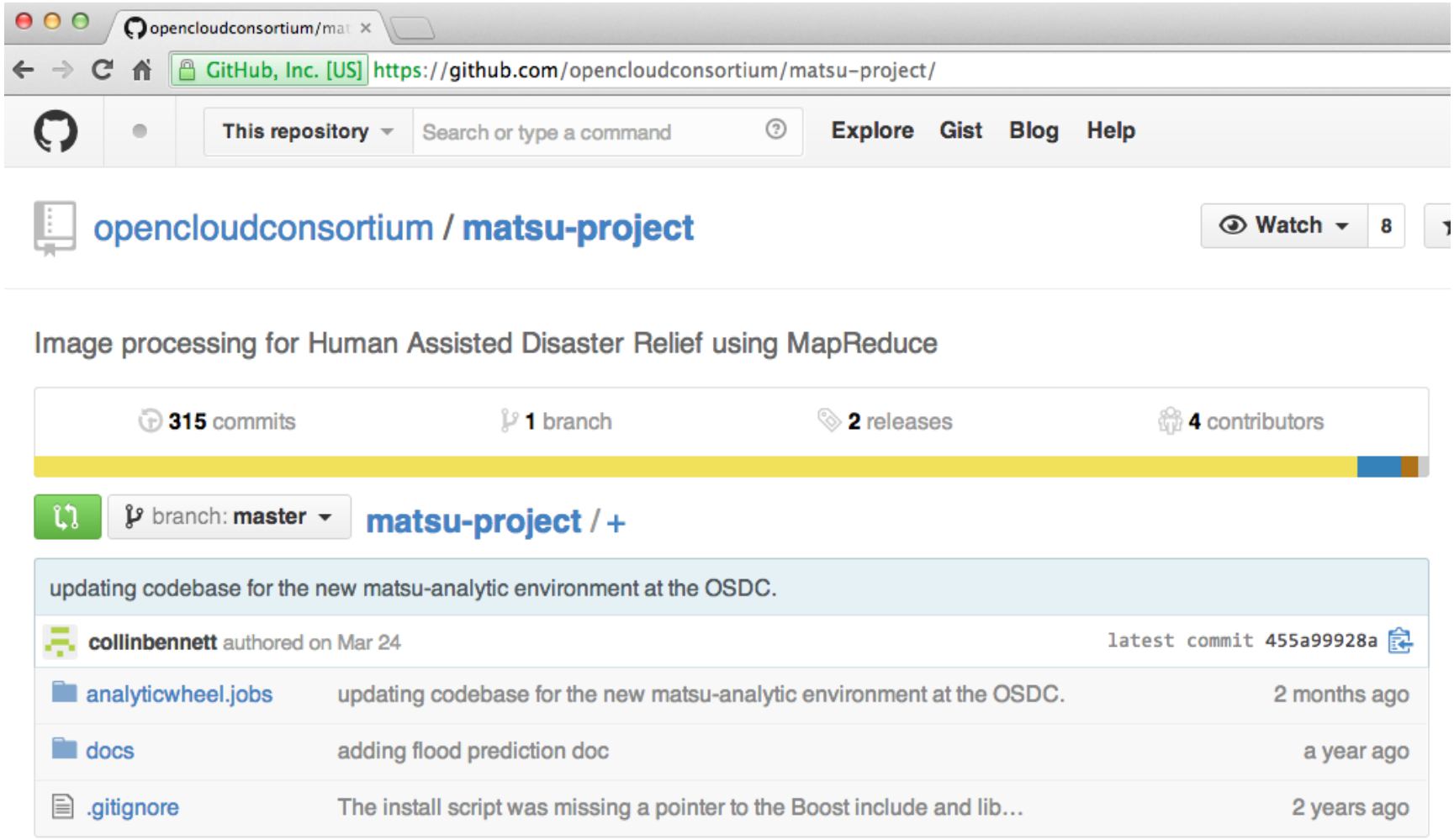
Collection Date	2014-02-25 (day 056)
Analysis Date	Thu Apr 17 01:45:31 2014
Analytic Environment	
Analytic	Contours-2013-12-r4
Noise Correction Enabled	True
Summary Stats	ss-2013-12-r1
Data Ingest	populateHDFS-2013-11-r1
Report Format	reportContoursR4
Hyperspectral Image	
Image	EO1H1330512014056110KF_HYP_L1G
Number of Bands	242



Contour ID	Cluster Score	Contour Score	lat,lng	Area (Pixels)	Area (Meters)	color	Spectral Signature
C2-33051-0KF	1000	0.4176	93.8552204394,12.2723263195	2.1740	137.7266	COLOR	wavelengths
C2-33051-0KF	1000	0.4510	93.8585620292,12.2742192321	103.3753	6549.0126	COLOR	wavelengths

Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*

Matsu Wheel is open source

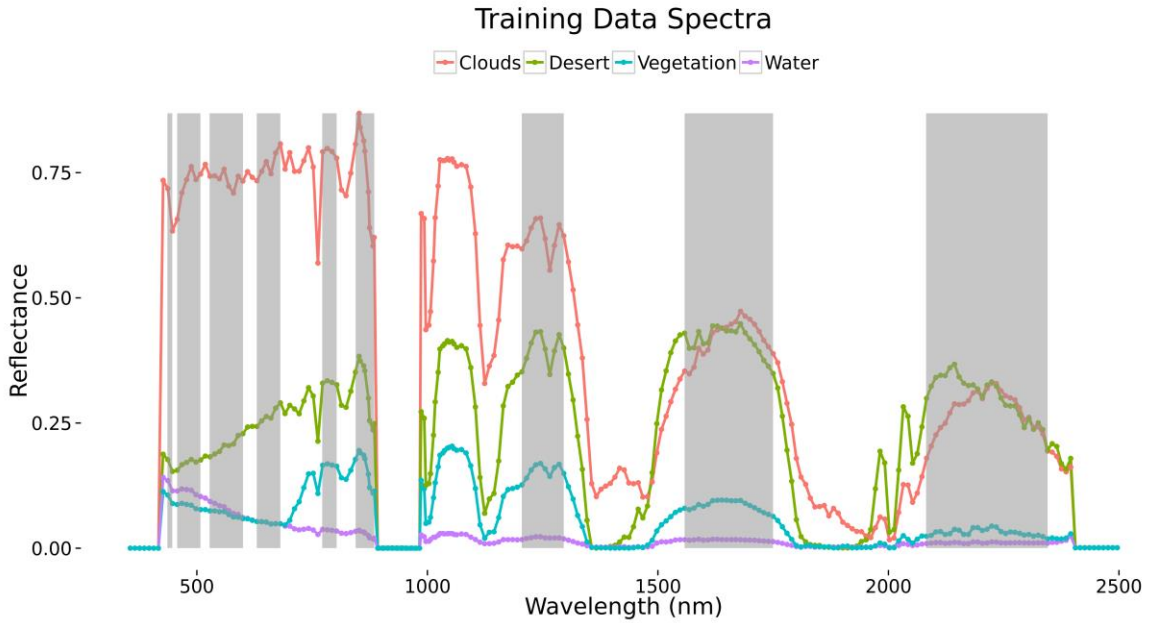


github.com/opencloudconsortium/matsu-project

Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*

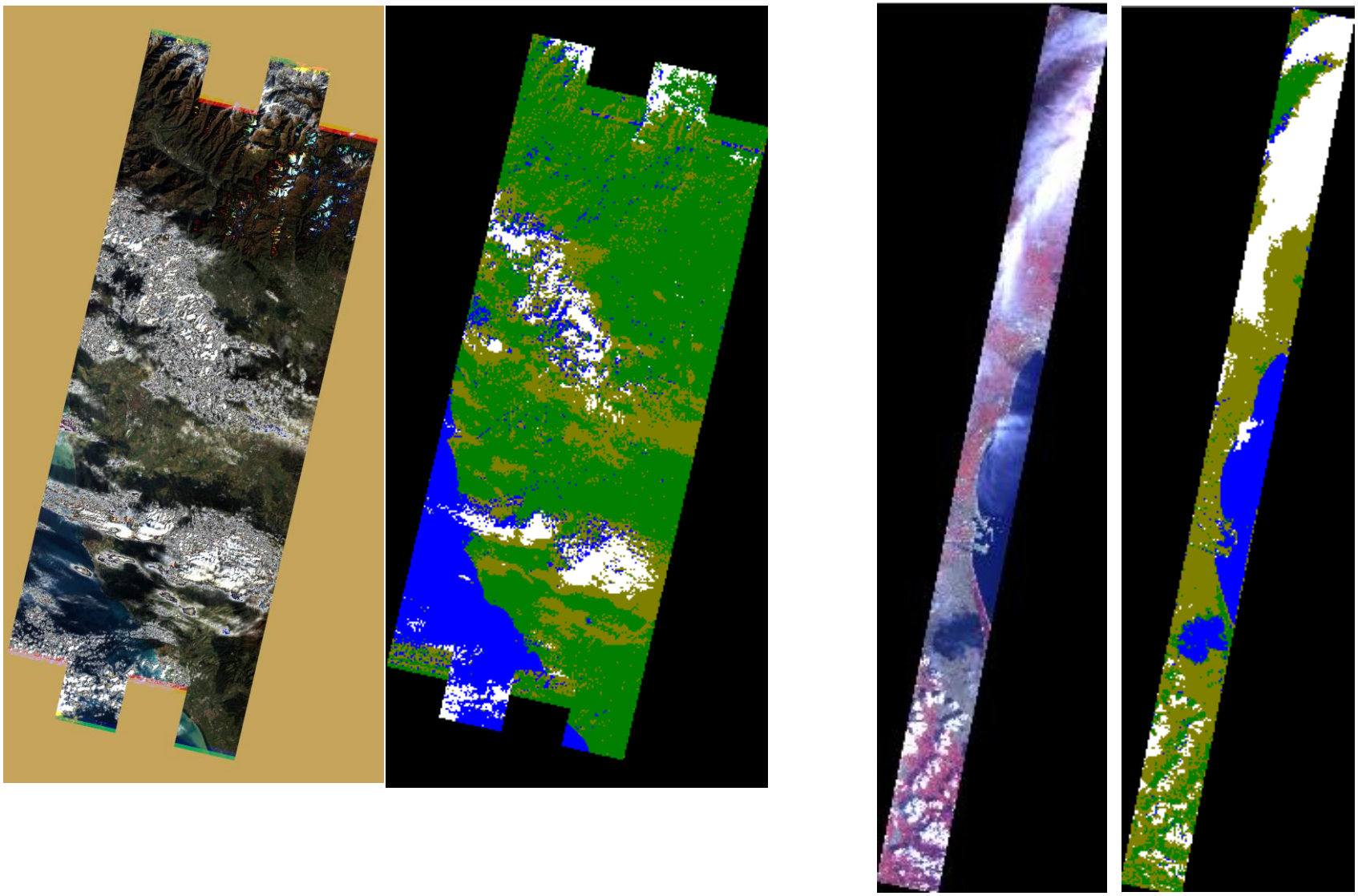
New wheel analytic (beta): Support Vector Machine (SVM) classifier

- A supervised machine learning classification algorithm
- Train the classifier by hand classifying areas in a set of training images
- Beta classifier has 4 classes: clouds, dry land, vegetation, water



Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*

New wheel analytic (beta): Support Vector Machine (SVM) classifier



Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*

Continuing work

- SVM classifier adapt regionally to geographic area (classes depend on geography)
- Incorporate SVM classifier into Matsu Wheel
- Additional wheel analytics
- Web Map Service and tiling using Geoserver
- Add additional data to the Wheel

What you can do

- Make your data available to Project Matsu
- Port your analysis tools and applications
- Use the Matsu cloud to facilitate making discoveries that require integrating multiple large datasets
- Contribute a Wheel analytic

Maria Patterson (mtpatter@uchicago.edu) *Center for Data Intensive Science, University of Chicago*